Modelling the Evolution of the Russia-Ukraine Conflict

36320 , 29921, 35190

April 26, 2022

Abstract

After the onset of the COVID-19 Pandemic, human civilisation globally has gone through a lot of emotional turmoil in the last three years as most of us have either lost our loved ones or our livelihoods. Moreover, just when people thought that things were returning to normal, Russia attacked Ukraine leading to more anguish in the minds of people worldwide. Thus, in order to decipher these emotions and developments properly, we conducted a study on the present war. Our study initially deals with the evaluation of people's feelings, as well as identifying key influential individuals using Topic Modeling, Sentiment Analysis and Graph Network Analysis, respectively. This is then followed by building a simulation study with the help of streaming process systems that will provide guidance to civilians about any treacherous events occurring in their vicinity in real-time, thereby cautioning them in split-seconds of potential threat to life and property, as well giving an unbiased news reporting of the ongoing happenings of the war. After the completion of our study, we came to the conclusion that the war has caused devastation in the minds of people as well as destruction, as critical topics identified within the tweets had to do with people seeking immediate help or bombing concerns and threats within Ukraine. This desperation of civilians suffering from the effect of war has undoubtedly led to people worldwide, inclusive of prominent world leaders, to turn against it, as identified from the sentiment analysis and graph data processing which was conducted. Finally, the streaming system developed, highlighted the fact that many twitter users are very opinionated with respect to the war, leading to difficulties in identifying the absolute true facts concerning its developments.

Keywords: Streaming Processing Systems, Graph Data Processing, Sentiment Analysis, Topic Modelling, Latent Dirichlet Analysis (LDA), PySpark, Distributed Computing, Big Data

1. Introduction

As a starting point to our study, we would like to give a brief overview of the history of the two countries and the plausible reasons for the ongoing war, to gain a better insight into its origins, thereby providing the reader some context about the ongoing situation. Russia and Ukraine share a long common history, though the former dominated the latter in the days of the Russian Empire and Soviet Union. Russian and Ukrainian families share a longstanding level of kinship, even though the Soviets took it further with enforced codes of brotherhood between the two (Hosking, 1997). Tensions dominated these relations between Moscow and Kiev since Ukraine's 2004 Orange Revolution (Zhurzhenko et al., 2011) marked its realignment in the direction of Europe which eventually led to the Russia-Ukraine War. This has been an ongoing war between Russia (together with pro-Russian separatist forces) and Ukraine since the Ukrainian Revolution of Dignity (Johnson, 2020). This war was initially focused on the status of Crimea and parts of the Donbas, internationally recognised as part of Ukraine (Robinson, 2016). The preliminary years of the conflict included the Russian annexation of Crimea and the war in Donbas between Ukraine and Russian-backed separatists, as well as naval incidents, cyberwarfare, and political tensions (Gutsul and Khrul, 2017). Following a Russian military build-up on the Russia–Ukraine border from late 2021, the conflict expanded significantly when Russia launched a full-scale invasion of Ukraine on 24 February 2022.

Moving on, if one has to describe the last few years in two words then 'Uncertainty' and 'Hopelessness' will top the charts. Thus, in order to reduce these anxieties from the minds of numerous individuals we conducted a study where we first analyze three different types of emotions *(Positive, Neutral* and *Negative)* of a person towards the war and then develop a system that will guide such individuals to be aware of any news or threatening events in the future.

The first part stated above is achieved with the help of three algorithms: Topic Modeling, Sentiment Analysis and Graph Network Analysis. As per Schmiedel et al. (2019), topic modeling represents a novel tool for analyzing large collections of qualitative data in a scalable and reproducible manner. For the scope of our study we have used it to illustrate critical topics from the perspective of the ongoing war. The paper then focuses on the second procedure that is Sentiment Analysis, to assess the overall sentiment of the world and also country specific opinion of the battle in progress, over time. After this, the opinion of top influencers globally and country-wise is analysed, and we identify whether there is any relation between the views of such people, with the people of their country and of the world at large, using Graph Network Analysis.

Coming to the realisation that the war has caused major suffering, as well as the fact that the tweets are biased by the strong opinions of people globally, gave us the motivation to build a streaming system, designed to identify unbiased non-opinionated tweets. These could act as a reference to people, to monitor the happenings of the war without news being clouded by people's individual judgement and opinion. The built streaming process thereby acts as a news reporting system, built on Twitter data obtained by the Twitter API. For instance, if there is an evacuation happening in the neighbouring areas of Kiev, then with the help of keywords relating to the war, people will get intimated about the on-going/upcoming evacuation drives and thereby this will help people to plan in a judicious manner. This will not only give them timely information but also help in reducing the burden that most individuals are facing at present, as they would be able to relocate to safer places as quickly as possible.

Along with this report, the code and images relating to the analysis can be seen viewed in the attached Jupyter Python Notebook called *"Modelling the Evolution of the Russia-Ukraine Conflict"*. The recommendation is to first read the report and then the notebook.

2. Data

In this study, one data base is employed. The reference data base can be found in `https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows`. This data set contains tweets monitoring the current ongoing Ukraine-Russia conflict and gets updated every day between 1 am to 3am UTC. The data studied was extracted from Twitter. As a popular social networking site, it is a source that contributes huge data which possesses value beyond social and commercial interests. Twitter users express their feelings or information regarding events, incidents, health or anything in their 140 character restricted short messages, termed tweets (Kwak et al., 2010). In-order to meet the objective of our project in the stipulated time frame, we fixed the end date as 17^{th} April 2022 for our study. This data-set includes close to 23.25 million tweets which amounts to 4.92 GB of zipped data.

Due to the very large scale of our data, we conducted our analysis in Google Cloud Platform. This enabled us to take advantage of both Google Cloud storage to store our data and Google Cloud Dataproc, which offers Apache Spark and Hadoop services for big data processing (Moreno et al., 2013). To load in and analyze the unzipped data, the configuration of the cluster had to be set to accommodate for its large magnitude. Specifically, spark executor, driver memory, and max results size, as well as yarn scheduler maximum allocation and node-manager resource memory, had to be significantly increased to avoid running out of Java heap space and the death of kernels (Sahin, 2019). Following the setup of the cluster and bucket used for storage, SSH was used to connect to the master node, where we used Kaggle's API and the commands stated in the Jupyter Notebook, to access, download and copy the data into the bucket. Then we started our exploration using PySpark. For the scope of our study, one should note that only the tweets that were in written in English have been considered, as it would have been cumbersome to comprehend the tweets written in multiple languages. Further, we have only taken into account the following countries; Ukraine, China, United States of America (USA/US), North Atlantic Treaty Organization (NATO), Non North Atlantic Treaty Organization (Non-NATO), The United Kingdom (UK), India and Belarus as after conducting initial checks on our data, these areas were underscored as the ones making the maximum number of tweets, as well as the ones most directly involved in the war, either on Ukraine or Russia's side. Please bear in mind that Russian tweets are not present in the data-set as a result of Twitter restrictions which possibly is a result of the falling down of Twitter restrictions on Russia, set these past months (Bonifacic).

The data-set had the following columns *userid, username, acctdesc, location, following, followers, totaltweets, usercreatedts, tweetid, tweetcreatedts, retweetcount, text, hashtags, language, coordinates, favoritecount* and *tweetcreatedts.* However, for the scope of our study we have only considered the below mentioned columns:

- *Username* - The name of the specific user who has tweeted.
- *Location* - Place where the user belongs or from where the user tweeted the said tweet.
- *Text* - The content that the user has stated/typed in the tweet.

- *Language* - The form of written speech that was used by the user.
- *Following* - The number of people that particular user is following.
- *Followers* - The total number of followers that particular user has. It also tells about the reach of a particular user.
- *Retweetcount* - This is the number of times a particular tweet given by one user has been shared by other users on the twitter platform.
- *Favoritecount* - The number of tweets that given user has marked as favorite.
- *Tweetcreatedts* - The date and time when the tweet was created.

For the streaming processing system only, the Twitter API was used to obtain tweets concerning the war in real-time.

3. Methodology and Metrics

Due to the fast growing world of data we live in and the subsequent advancement in technology, a transition from Traditional Data such as paper surveys conducted by most government organisations across the globe to Big Data Processing Systems, was pertinent (Nair and Shetty, 2015). Big Data, refers to data that is so large, fast or complex that it's tedious to process using traditional methods (Elgendy and Elragal, 2014). This has therefore led to the development of various computing frameworks such as Apache Spark, Apache Hadoop and Kafka for big data processing (Nair and Shetty, 2015). As we start discussing the methodology of our report, we want to first explain the primary reasons as to why we opted for Apache Spark as the fundamental tool to conduct our analysis of Big Data. Spark is an open-source with an efficient computational speed from the realm of big data distributed processing platforms (Nair and Shetty, 2015). It has similar principles as Hadoop, but the place where it overpowers Hadoop is graph based algorithm tasks which is one of the essential components of our analysis (Elzayady et al., 2018). Furthermore, it does not depend on other platforms in order to perform streaming processes (Nair and Shetty, 2015).

Another drawback of working with Hadoop, is that it is very confined in its functions, i.e. it has only two components ”Map” and ”Reduce”, thereby making the task of dealing with Big Data more cumbersome from the context of Machine Learning (Elzayady et al., 2018; Nair and Shetty, 2015). Apache Spark has the Spark MLLib which is defined as a distributed machine learning framework on top of Spark Core (Nair and Shetty, 2015). This library is especially curated to carry out a plethora of machine learning models in the Spark framework, something that the Hadoop framework lacks. Further, the latter only operates in one programming language that is Java as opposed to the former where developers can also use Scala, Python and R (Chambers and Zaharia, 2018; Nair and Shetty, 2015). The existence of such multiple programming languages, the multiple ecosystems that it contains (MLLib, GraphX, Spark SQL and Spark Streaming), the higher memory storage and the fact that Spark can be scaled in the context of the number of nodes that a user needs, are beneficial to the application of the chosen methodologies (Duan et al., 2016; Nair and Shetty, 2015).

Moreover, Spark also consists of Resilient Distributed Data-sets (RDDs). RDDs are ”fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators” (Zaharia et al., 2012).

This component was extremely essential from the context of Topic Modeling where one has to segregate the data into logical parts in order to correctly classify each topic on the basis of its distribution in the data. The RDD can be constructed in two manners, one can be by parallelizing an existing collection in your Spark Context driver program (Ramírez-Gallego et al., 2018). The other way is by referencing a data set in an external storage system that can be HDFS, or any other source which has Hadoop file format. For the scope of our study, we have used the first approach in order to implement are algorithms.

3.1 Topic Modeling

Topic Modeling is an unsupervised machine learning algorithm, that analyzes text data to cluster words for a set of documents. It does not require a predefined list of words or training data that's been previously classified by humans. The procedure of topic modeling consists mainly of ”words”, ”documents” *(in the context of lay-man splitting the topic into unique words)*, and ”corpora” *(a dictionary that is split up on the basis of specific words)*(Negara et al., 2019).

Imagine one has a set of predefined words, these words then have n number of subsets in the form of "documents" within it. These documents are then segregated into m number of sub-sets also known as "corpora" on the basis of varied themes available in the English dictionary. Thus each document in the corpus contains its own proportions of the words discussed according to the meaning of the word (Cortés Hinojosa, 2016). This proportion guides us in interpreting the topics in descending order i.e. the one with the highest weightage indicates the most tweeted topic.

Another name for the terminology explained above is Latent Dirichlet Allocation (LDA), a generative probalisitic model of a corpus. In general, LDA works with the initialization of discrete documents and several parameters, to generate results in the form of a model consisting of term weights that can be normalized on the basis of probability (Blei et al., 2003).

Mathematically, LDA makes the following assumptions when generating each document w in the corpus D (Blei et al., 2003).

1. $N \sim Poisson(\xi)$
2. $\theta \sim Dir(\alpha)$
3. For each of the N words w_n: Choose a topic $z_n \sim Multinomial(\theta)$ and choose a word w_n from $p(w_n|z_n, \beta)$

Thus the probability approach followed by LDA is two fold, first is the probability of the specific words a topic consists from the context of English vocabulary, and then the probability of the occurrence of that topic or similar topics from the context of data-set in hand. From the point of view of our study, the probability measure here is considered in the context of term-weights (Jelodar et al., 2019). LDA conditional probabilities given parameters α and β can be defined as follows (Blei et al., 2003):

The Joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

The Marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta.$$

The probability of a corpus:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

For topic models such as LDA, a commonly used indicator is perplexity, an accuracy metric of a probability model and to calculate the held out log-likelihood, where a lower perplexity indicates a better prediction (Huang et al., 2017). we first train an LDA model on a portion of the data. Then, the model is evaluated using the held-out data. This routine is repeated for models with different numbers of topics, so that it becomes clear which amount leads to the lowest perplexity.

$$Perplexity = -exp\left(-\frac{\sum_{d=1}^{m} \log\left(p(W_d)\right)}{\sum_{d=1}^{m} N_d}\right)$$

Here, m is for the total number of documents and N_d is the number of clusters for the corpus, for each document d.

3.2 Sentiment Analysis

Sentiment analysis, is one of the most popular social media text analysis tools that helps in assessing the views of people on numerous topics across the world. From the point of view of our study, it helps us in interpreting whether a tweet given by a user is in favour, neutral or against the war. This process of division is carried out using a natural language processing (NLP) algorithm with the help of the NLTK package (Bird et al., 2008), which allows us to systematically measure, extract, identify, and evaluate effective opinions of users on the twitter platform (Hemalatha et al., 2013).

An analysis of people's emotions can be classified in a variety of ways. The first is text classification also known as text categorization and is the primary process in sentiment analysis (Govindarajan, 2013). It entails categorising texts into organised groups and calculating sentiment analysis based on the number of positive and negative word occurrences in each document, in this case each tweet. Each tweet received either a positive, neutral or negative rating (Khoo and Johnkhan, 2018).

This methodology is used, in order to first analyze the overall opinion of the world and then drill it down further to analyze the country specific opinions. Taking inspiration from the following literature Shakhov et al. (2020), we calculated the sentiment score using the module Sentiment Intensity Analyzer from VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Library. This method is preferred to other algorithms in libraries such as LIWC (Meduru et al., 2017). VADER is a lexicon and rule-based feeling analysis instrument that is explicitly sensitive to suppositions communicated in web-based media (Khanam and Sharma, 2021).

The lexicon approach has many advantages with respect to other sentiment analysis methods as aside from words, its lexical dictionary contains phrases, sentiment-laden acronyms and emoticons (Aljedaani et al., 2021). A score of polarity and intensity is assigned to each of the lexical features (Tymann et al., 2019). These scores are within the interval $[-4, 4]$, the more positive indicating a Positive sentiment analysis (Elbagir and Yang, 2019). The dictionary of the VADER approach contains around 7,500 sentiment features, which makes it a very rigorous/consistent approach (Aljedaani et al., 2021). The features not contained in the dictionary will be considered "neutral"(Tymann et al., 2019).

Besides having a very rich dictionary, the method has some heuristic rules to deal with the punctuation, capitalization, degree modifiers (booster words), contrastive conjunctions and negation statements (Bonta and Janardhan, 2019). As followed by Newman and Joyner (2018), once the VADER score has determined the sentiment features, these scores are summed and normalized to obtain the Compound Score (C), which is defined as follows (Gupta et al., 2020):

$$C = \frac{x}{\sqrt{x^2 + \alpha}}$$

The Compound Score ranges between the values -1 and 1 and the variable α is fixed to be 15, which approximates the maximum expected value of the sum of the scores, x (Newman and Joyner, 2018). Tweets having a positive Compound Score are considered to be Positive while the ones with a negative score are Negative (Elbagir and Yang, 2019). If the values of the Compound Score are close to 0, then the tweets are considered neutral (Elbagir and Yang, 2019).

3.3 Graph Network Analysis

A graph database can store data entities and relationships using simple concepts derived from mathematical graph theory. Nodes represent the graph data entities. Relationships are implemented as edges which connect the nodes. Properties are attributes that are associated with data entities and relationships, and are expressed as key-value pairs, e.g. (username: Alex). One common use of graph databases is the representation of social networks. Examples of graph database systems are Neo4j and Apache Spark. In the context of social networks, graph mining involves the analysis of links between social media users. Most researchers have reported that Twitter graph mining has been used to investigate interesting problems such as measuring user influence and the dynamics of popularity, community discovery, and community formation in social networks.

For our use case, the PySpark package Graphframes is employed, by which one is able to formulate queries, conduct motif finding and derive metrics such as Page Rank, for dataframe-based graphs on Apache Spark (Dave et al., 2016).

Page Rank, an iterative technique originally proposed by Page et al. (1999) to assess the importance of every web page based on the graph of the web, can be applied to the case of social network analysis, to rank the importance of each node i.e user, based on the in-degree of that node, and Pagerank and propensity or in other words out-degree of the other nodes that point to that node. It can be defined as (Franceschet, 2011; Page et al., 1999; Priyanta et al., 2019; Sharp et al., 2020):

$$PR(p) = \frac{1-d}{N} + d \sum_{i} \frac{PR(i)}{C(i)}$$

PR(i) - Page Rank of node i where i points to node p.

d - Damping Factor which takes values between 0 and 1. (Usually set to 0.85)

C(i) - Number of outgoing links/edges of node i.

N - Total number of nodes.

Given the nature of our data, which lacked information on the specific names of *followers* and *following* of each user due to privacy issues, to link the users between them, users were considered to be vertices, and the mentions were used as directed edges. As stated in Twitter's documentation, a mention is a Tweet that contains another person's username anywhere in the body of the Tweet, whereby each username is denoted by the symbol '@' followed by the name. Therefore, if a user mentioned another user within the text of his tweet, the two nodes representing each user are connected by an edge with source being the user who wrote the tweet and destination the mentioned user. As is done in previous literature such as Cha et al. (2010), identifying the most mentioned people and their importance using PageRank, serves as an initial indication of a user's relevance within the community of people tweeting about the war.

Further to the number of times a user is mentioned, two other metrics are used to draw insights into how influential a user is on the network. As mentioned in Saleiro and Soares (2016), one of the notions of popularity associated with entities, consists of the number of retweets on Twitter. Moreover, the number of followers a user has is a direct indication of the size of the audience of that user (Saleiro and Soares, 2016), which again has an impact on the level of influence that user holds.

As a result three metrics are defined to quantify a user's influence and importance:

Relevance score - *The number of times a user is mentioned by other users*

Popularity score - *The average number of retweets*

Reach score - *The difference between the number of followers of a user and the number of following of a user*

To make each score comparable given the differing scales of each metric, these are log transformed and normalised, to all have values ranging from 0 to 1. They are then combined together to create the *Total Influence Score.*

Total Influence score - *The sum of the transformed Relevance, Popularity and Reach scores. (Has a maximum value of 3)*

3.4 Streaming Processing Systems

Streaming processing systems are a big data tool which focuses on retrieving and processing (through filters) in real-time, sequences of data elements (Nair and Shetty, 2015). This is advantageous if the information to be obtained is needed immediately and cannot wait until the data is stored in large batches to be processed (Nair and Shetty, 2015). Furthermore, these systems reduce latency and allow scalability in order to process large volumes of data (Samosir et al., 2016). Although there are many streaming processing systems like Kafka, Apache Storm or Amazon Kinesis (Chintapalli et al., 2016), in this paper the focus will be on Spark streaming. In addition to the characteristics of reduced latency and scalability of streaming processes, PySpark is fault tolerant and has a fast recovery after failures (Nair and Shetty, 2015).

The streaming process in PySpark can be done in two different ways: non-structured streaming and structured streaming (Yadranjiaghdam et al., 2017). Our study will implement structured streaming which is built on the Spark SQL engine and uses the DataFrame API (Zaki et al., 2020). The input stream can come from different sources, it is then processed by using different operators and then outputted to different systems (Zaki et al., 2020), as shown in Figure 1. The structured streaming PySpark process can be summarised in Figure 2, which depicts that the data is inputted in batches, then processed by a query and then appended as new rows of an unbounded table (Armbrust et al., 2018; Zaki et al., 2020). These rows are added to the table after specific time intervals and these intervals can be specified by the variable trigger (Yadranjiaghdam et al., 2017). There are different output modes that can be chosen in order to specify how the result is updated after each time interval (Armbrust et al., 2018), however in our case only new rows will be added to the result table after each trigger, hence we will use "Append" as the output mode. There are also many sources to input the data that will be processed by the streaming system, namely file sources, Kafka source, socket source and rate source (Yadranjiaghdam et al., 2017; Zaki et al., 2020). In this paper the data will be inputted from the Twitter API and hence a socket source will be used. Structure spark streaming also has other features like the option to do window operations and handle late data by the use of watermarking (Armbrust et al., 2018).

In the streaming process a sentiment analysis algorithm will be applied, TextBlob. This NLP method employs the NLTK library (Gujjar and Kumar, 2021; Loria et al., 2018). The choice of this algorithm is convenient as

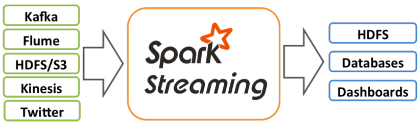


Figure 1: PySpark Streaming Formulation

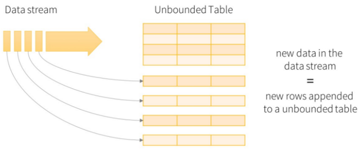

Data stream
Unbounded Table
new data in the data stream
=
new rows appended to a unbounded table

Figure 2: Structured streaming PySpark Process

it has less computational complexity than other sentiment analysis methods (Gujjar and Kumar, 2021). The tweets are passed to the algorithm and it gives a subjectivity and a polarity score for each one (Loria et al., 2018). Polarity and subjectivity are defined to be in the intervals $[-1, 1]$ and $[0, 1]$, respectively (Chaudhri et al., 2021; Loria et al., 2018). It is a lexicon-based approach, hence it assigns a score (based on a given sentiment dictionary) to each word or bag of words and then it takes the average in order to find the polarity and subjectivity scores of each of the sentences (Shekhawat, 2019). The subjectivity of each word is computed based on the intensity, which measures how much influence a word has in ”modifying” the next word (Shekhawat, 2019). The higher the subjectivity score the more biased is the tweet (Chaudhri et al., 2021; Gujjar and Kumar, 2021). On the other hand, the polarity determines how Positive (closer to 1) or Negative (closer to -1) is a tweet (Chaudhri et al., 2021; Gujjar and Kumar, 2021).

4. Evaluation and Implementation

4.1 Topic Modeling

4.1.1 Process

The first step in our study was to understand the top 20 most discussed topics from the perspective of the war. Here, the topics were classified on the basis of the Text column of the data-set under study. We started the process of Topic Modeling by selecting only four columns *Username*, *Text*, *Language* and *Location*.

This was then followed by some data-preprocessing techniques using RDDs, when we converted the text column into an RDD. Then we used libraries such as Stopwords, String Punctuation and Lemmatizer to get rid of extra spaces, characters and words that are not significant from the scope of this study. For this removal of Stopwords we set a threshold to the count of each word of 3000, as after running the code multiple times we discovered that at this level, the number of words that are not related reduces substantively. (Saif et al., 2014)

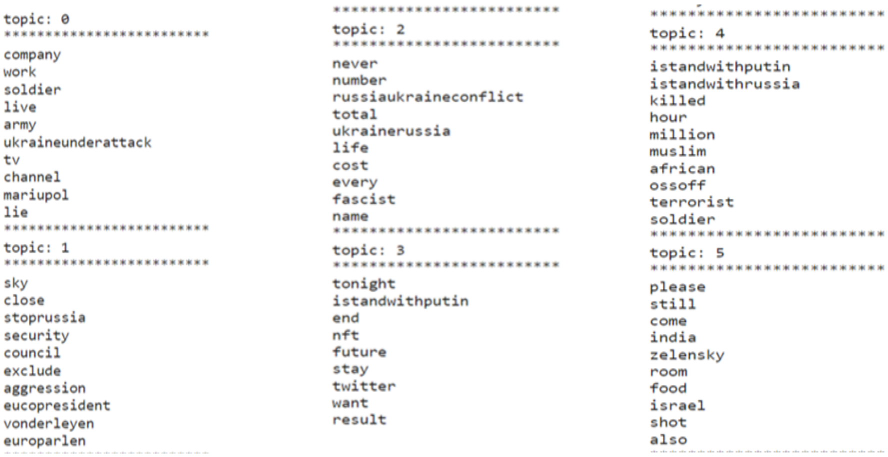
Moving on to the vocabulary check, the main goal for this was to understand the number of distinct and logical words that are present in our data-set. After conducting this we identified close to 26815 unique words. This in our opinion was sufficient enough for us to further conduct an LDA analysis and calculate the perplexity.

As discussed in Section 3.1, one has to first train an LDA model on the given data-set and then evaluation is carried out on the basis of perplexity. Thus, the LDA model was trained on for five iterations, followed by which the perplexity was calculated. Having trained an LDA model for different topic sizes, we identified that choosing 20 topics led to the lowest perplexity results in particular we got a lower bound on the log likelihood of the entire corpus score of approximately-2270915 and an upper bound score of the perplexity is 9.5, which indicates that the model is good enough to further analyze the topics with high weightages.

The last few steps are to describe the topics, estimate the term weights and then sort the topics on the basis of the one having the highest weights to the one having the lowest ones. The results of the same are discussed in the next section.

4.1.2 Results

The above figure (Figure 4) depicts the Top 6 topics being discussed by the plethora of users on twitter. It is observed that, either the people are asking for help, as words mentioned in topic 5 include *”please, still, come, India, zelensky, food”*, suggesting that people in Ukraine may be requesting the people in India to send some food-aid. Moreover, in topic 1, the words *security, council, eucopresident, europarlen* suggest that people are pleading for European Intervention in the war to stop the *”fascist”*, *”aggression”* and *”terror”* as seen in the remaining topics. On the other hand, topic 4 suggests that there are people in favour of the war, as portrayed from the words *”istandwithputin, istandwithrussia”*. These strong polarized responses, highlighted in this section, further gave us the urge to perform Sentiment Analysis, in-order to understand the topics in greater depth and get some direction towards the overall view of the people.



```
topic: 0
************************
company
work
soldier
live
army
ukraineunderattack
tv
channel
mariupol
lie
************************
topic: 1
************************
sky
close
stoprussia
security
council
exclude
aggression
eucopresident
vonderleyen
europarlen

************************
topic: 2
************************
never
number
russiaukraineconflict
total
ukrainerussia
life
cost
every
fascist
name
************************
topic: 3
************************
tonight
istandwithputin
end
nft
future
stay
twitter
want
result

************************
topic: 4
************************
istandwithputin
istandwithrussia
killed
hour
million
muslim
african
ossoff
terrorist
soldier
************************
topic: 5
************************
please
still
come
india
zelensky
room
food
israel
shot
also
```

Figure 3: Top 6 Topics being discussed

4.2 Sentiment Analysis

4.2.1 Process

Topic Modeling gave us the head start to implement the second approach of our study, the Sentiment Analysis. It was a stepping stone to delve deeper into the opinion of the general population of the world and to assess if there was any relation between the world leader of a specific country and the local population of that country. The fundamental motivation for this method was to see if there is actually a sync between the people and their leaders.

This process followed a four-fold approach. First, to clean the tweet in order to get rid of all the superficial terms and spaces in our data-set, as is done in Topic Analysis. Second, to convert it into a Tokenizer. Tokenization breaks the raw text into words called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words (Chakravarthy, 2020). The third step was to remove the stop words and the final step was to calculate the sentiment score. The process is displayed in Figure 4.

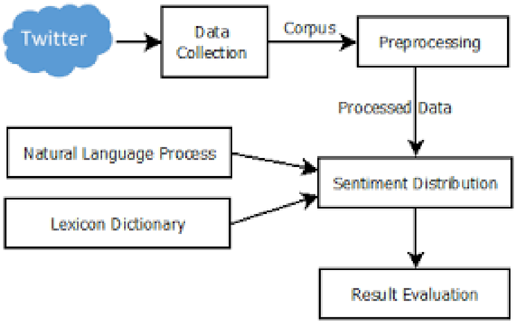


Figure 4: Sentiment Analysis Procedure

The sentiment score is calculated for the entire world population and for the countries mentioned in Section 2 of this report. After carrying out Sentiment Analysis, we also created wordclouds for our data-set. A wordcloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is (Papageorgiou, 2021). For the scope of our study, wordclouds are created for frequent words in our tweets and for all the words in the dataset.

To give a small qualitative outlook of how people's sentiment towards the war has changed throughout its development, the single most retweeted tweet of the dates 2022-02-27, 2022-03-09, 2022-03-19, 2022-03-29, 2022-04-08 (every 10 days), was also reported and evaluated.

4.2.2 Results

This analysis helped in clearing out the confusion that aroused in Topic Modeling. After, carrying out the analysis, the overall sentiment score achieved is **-0.606** approximately for the entire population under study. Thereby suggesting that on a large scale, most of the people are at odds with the war. As we further drill down country-wise we see that out of all the countries taken into consideration, the only one that has a positive sentiment score is Belarus, thus suggesting the country is in favour of the war. This makes sense as Belarus is one of the nations that has been supporting Russia in the war (Ambrosio, 2006). Moreover, countries such as China and India have maintained a mostly neutral position, perhaps to maintain good relations with both the countries. The countries significantly against the war are the United Kingdom, United States of America and NATO, perhaps as a result of Russia's aggression of Ukraine entering the NATO (Wolff, 2015). The country-wise analysis also affirms the fact that as a whole most nations want peace and sanity for all. The specific country-wise scores are given in Figure 18 of this report.

Below attached is the word-cloud depicting the words that are most frequently used in the data-set (Figure 5). It is observed that anxiety prone and fearful words such as *”destroy”*, *”break”* and *”pistol”* are being highlighted along side positive and peace promoting words such as *”reparation”* and *”friend”* there-by suggesting that people want this hatred to come to an end, and promote peace and happiness for all.

When assessing sentiment across time, the preliminary results indicated that the tweets which were the most retweeted in the mentioned dates, all contained the same text, specifically:

.@ZelenskyyUa's tv address to the Russian (!) people might be the most moving speech that I've ever seen in my entire life. The whole world needs to see, understand and share this crucial Ukrainian message. #StandWithUkraine #Ukraine # #Russia # https://t.co/WoMOgqXTWX

This clearly portrays the fact that global support for Ukraine has remained unfaltering as the war has progressed. To try and obtain a more diverse and holistic view, unobscured by tweets only with this specific text, all tweets including this text were removed and the same process of finding the most retweeted tweet every 10 days was carried out. As seen from Figure 6, once again, all tweets are pro-Ukranian, relating to the difficulties civilians are facing and shedding light on the existence of alleged Russian propaganda.

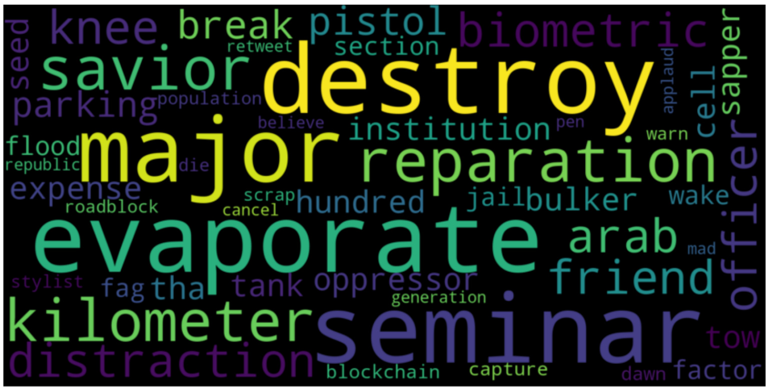
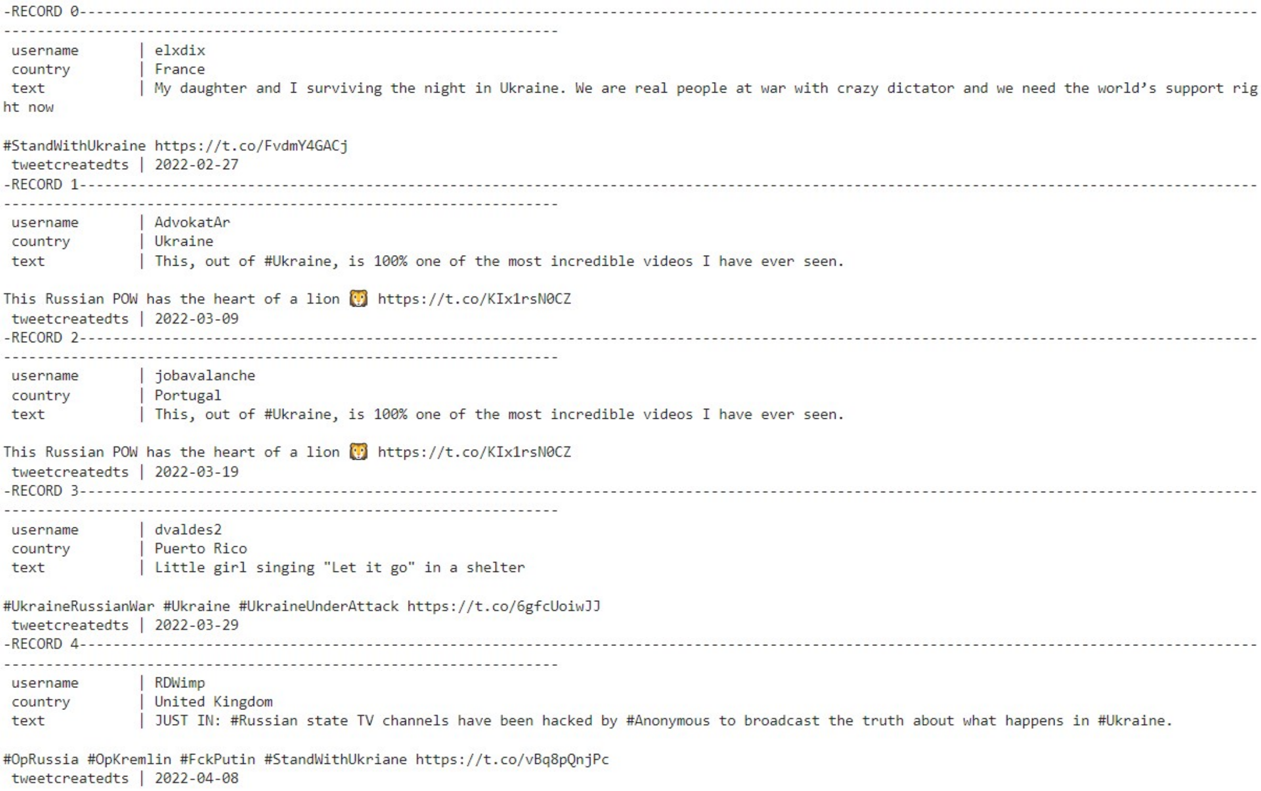


Figure 5: Wordcloud of the highest frequency words



```
-RECORD 0--------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------
 username       | elxdix
 country        | France
 text           | My daughter and I surviving the night in Ukraine. We are real people at war with crazy dictator and we need the world’s support rig
ht now

#StandWithUkraine https://t.co/FvdmY4GACj
 tweetcreatedts | 2022-02-27
-RECORD 1--------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------
 username       | AdvokatAr
 country        | Ukraine
 text           | This, out of #Ukraine, is 100% one of the most incredible videos I have ever seen.

This Russian POW has the heart of a lion 🦁 https://t.co/KIx1rsN0CZ
 tweetcreatedts | 2022-03-09
-RECORD 2--------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------
 username       | jobavalanche
 country        | Portugal
 text           | This, out of #Ukraine, is 100% one of the most incredible videos I have ever seen.

This Russian POW has the heart of a lion 🦁 https://t.co/KIx1rsN0CZ
 tweetcreatedts | 2022-03-19
-RECORD 3--------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------
 username       | dvaldes2
 country        | Puerto Rico
 text           | Little girl singing "Let it go" in a shelter

#UkraineRussianWar #Ukraine #UkraineUnderAttack https://t.co/6gfcUoiwJJ
 tweetcreatedts | 2022-03-29
-RECORD 4--------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------
 username       | RDWimp
 country        | United Kingdom
 text           | JUST IN: #Russian state TV channels have been hacked by #Anonymous to broadcast the truth about what happens in #Ukraine.

#OpRussia #OpKremlin #FckPutin #StandWithUkriane https://t.co/vBq8pQnjPc
 tweetcreatedts | 2022-04-08
```

Figure 6: Most retweeted tweet every ten days starting from 2022-02-07

4.3 Graph Network Analysis

The next step in our investigation is two-fold:

1. Finding most influential users when it comes to the war, globally and by country.

2. Assess the sentiment of influential users and compare it with the sentiment of the country they originate from.

The motivation behind this lies with the fact that social media influencers have persuasive power on their following, as by definition an influencer is ”someone who affects or changes the way that other people behave” (Brian J. Taillon, 2020).

As a result, identifying the most influential people with respect to the war, will give us a clear picture of who is holding the strings of social media and has the ability to sway their following, whether these are institutions, politicians, famous people or people who have never previously been in the limelight. Analysing the views of these influencers will also allow us to investigate whether such users, are in agreement with the rest of the public. It is often the case where the views of the few in power, are different from the views of the many. Therefore making a comparison of the sentiment of the key users and of the public country-wise, allows for an investigation of whether the people are represented on social media by those who adhere to the same views.

4.3.1 Process

Let’s first turn to identifying key users. As mentioned previously, as a first indication of user’s influence and relevance, graph networks are built using the mentions of people within the body of text of the tweets. The text was thus filtered to identify the mentioned users of each tweet and to collate a dataframe of user and *mentioned user*. Using user and *mentioned user* as nodes, and the predicate ‘mentioned’ as directed edges between them, an initial graphframe was built to enable the identification of the most mentioned users globally as well as their interconnectivity. The Pagerank algorithm was carried out to assess the centrality of each user, and this network was then visualised using the NetworkX package, where the top 10 users with the highest in-degree, which corresponds to the users with the most mentions, are clearly marked by their label and node size. The graph was also filtered by the country of the user writing the tweet and once again visualised to obtain an indication of the similarities and differences between the people being mentioned by users in specific countries.

]

Using the overall number of times each user was mentioned, we were able to formulate the aforementioned *relevance score*, which will ultimately be used to assess user influence. The remaining *popularity* and *reach scores* were calculated per user, to thus finally reach a finalised total score for each user as described in the previous section. The top 10 users with the highest Total score, overall and by country, were identified, and are displayed in the form of bar charts in the results section.

To now tackle the second objective of identifying the sentiment of the most influential people, sentiment analysis was performed only on the tweets of the top influencers overall, and an average sentiment of the tweets was calculated. This was repeated country-wise, where the average sentiment of the top influencers from each country was determined, allowing for direct comparison with the previously derived average sentiment of each country (Section 4.2).

4.3.2 Results

Based on the network of connections between the users overall (Figure 7) it is clearly evident that the most mentioned user by far is *POTUS* referring to the President of the United States, *Joe Biden*, with 1577232 mentions as seen in Figure 11, followed by *NATO*, *RTErdogan* which corresponds to Turkish President *Tayyip Erdogan*, and Ukrainian President *Volodymyr Zelensky*. The rest of the most mentioned users are also leading politicians in many countries, which comes as no surprise given the large pressure for them to be outspoken and take action with relation to the war. What is noteworthy however is that, although some more than others, all the top mentioned users have publicly frowned upon the ongoing war, and thus this may reflect the feelings of the people towards it as well. Ukrainians seem to be mentioning primarily European bodies as seen in Figure 8, perhaps as a cry out for support, whereas British people (Figure 9) mainly mention Russian embassies around the world. Given the stance of Britons, who have shown to be in support of Ukraine, it is reasonable that they are calling out Russian Twitter accounts in their plea for an end to the conflict.

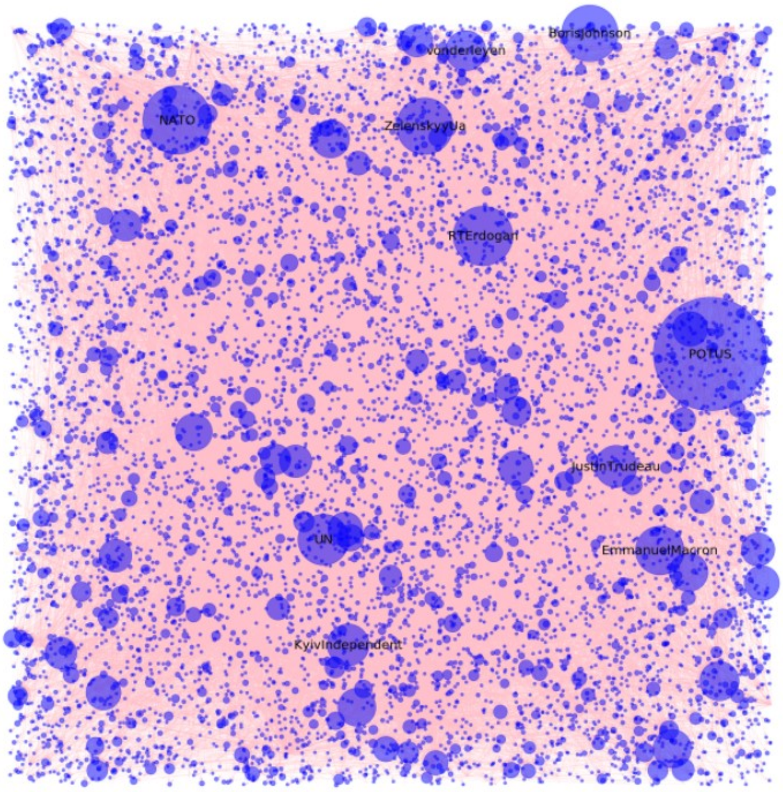


Figure 7: Graph network of mentions globally

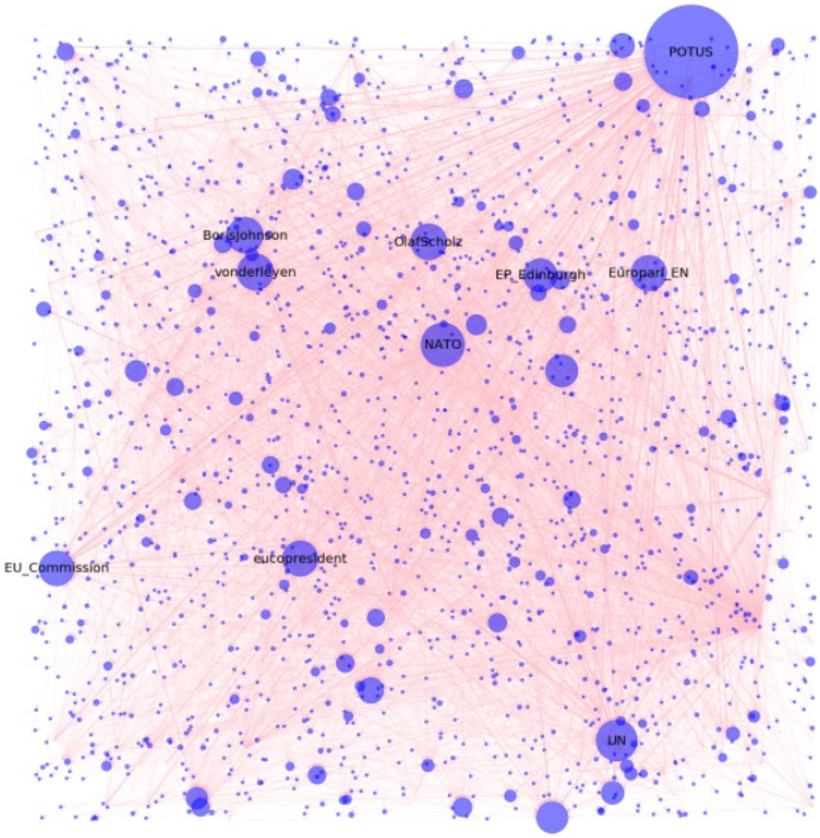


Figure 8: Graph network of mentions by people in Ukraine

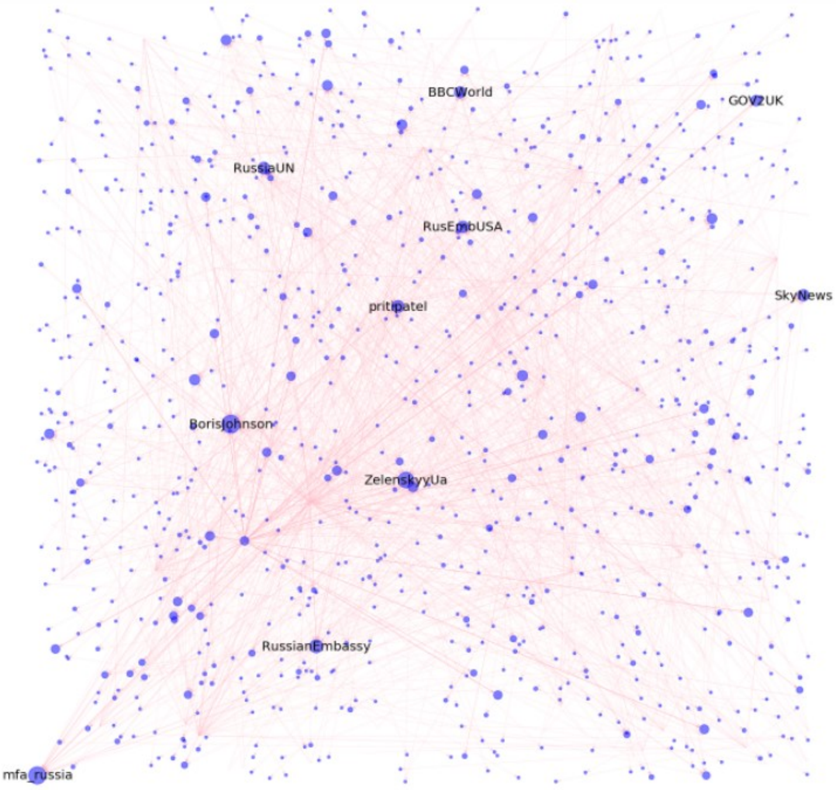
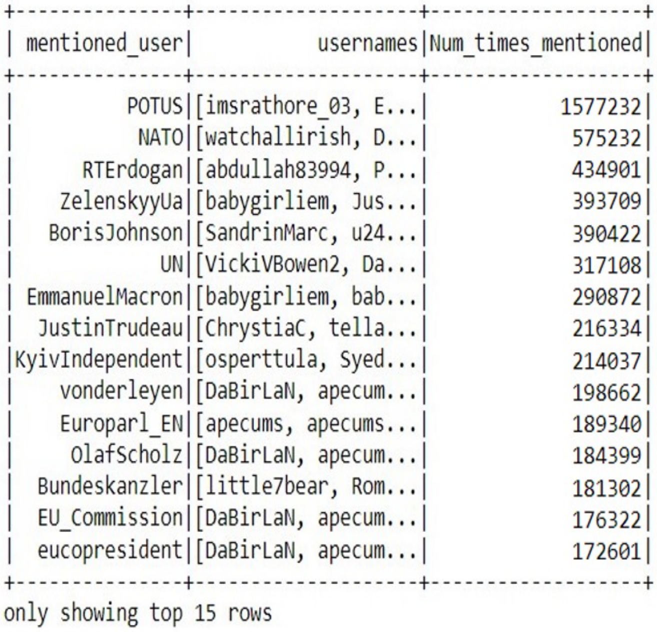
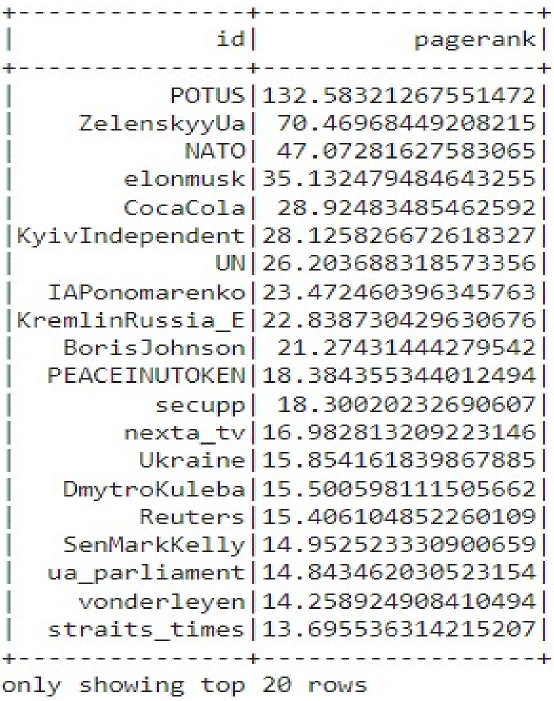


Figure 9: Graph network of mentions by people in the UK



```
+---------------+--------------------+-------------------+
| mentioned_user|           usernames|Num_times_mentioned|
+---------------+--------------------+-------------------+
|          POTUS|[imsrathore_03, E...|            1577232|
|           NATO|[watchallirish, D...|             575232|
|      RTErdogan|[abdullah83994, P...|             434901|
|    ZelenskyyUa|[babygirliem, Jus...|             393709|
|   BorisJohnson|[SandrinMarc, u24...|             390422|
|             UN|[VickiVBowen2, Da...|             317108|
| EmmanuelMacron|[babygirliem, bab...|             290872|
|  JustinTrudeau|[ChrystiaC, tella...|             216334|
|KyivIndependent|[osperttula, Syed...|             214037|
|    vonderleyen|[DaBirLaN, apecum...|             198662|
|    Europarl_EN|[apecums, apecums...|             189340|
|     OlafScholz|[DaBirLaN, apecum...|             184399|
|  Bundeskanzler|[little7bear, Rom...|             181302|
|  EU_Commission|[DaBirLaN, apecum...|             176322|
|  eucopresident|[DaBirLaN, apecum...|             172601|
+---------------+--------------------+-------------------+
only showing top 15 rows
```

Figure 10: Most mentioned users



```
+---------------+------------------+
|             id|          pagerank|
+---------------+------------------+
|          POTUS|132.58321267551472|
|    ZelenskyyUa| 70.46968449208215|
|           NATO| 47.07281627583065|
|       elonmusk| 35.13247948464325|
|       CocaCola| 28.92483485462592|
|KyivIndependent|28.125826672618327|
|             UN|26.203688318573356|
|  IAPonomarenko|23.472460396345763|
|KremlinRussia_E|22.838730429630676|
|   BorisJohnson| 21.27431444279542|
|  PEACEINUTOKEN|18.384355344012494|
|         secupp| 18.30020232690607|
|       nexta_tv|16.982813209223146|
|        Ukraine|15.854161839867885|
|   DmytroKuleba|15.500598111505662|
|        Reuters|15.406104852260109|
|   SenMarkKelly|14.952523330900659|
|  ua_parliament|14.843462030523154|
|    vonderleyen|14.258924908410494|
|  straits_times|13.695536314215207|
+---------------+------------------+
only showing top 20 rows
```

Figure 11: Users with the highest Page Rank score

Looking now at the page rank (Figure 10), similar users to the ones previously identified can be seen to have the highest page ranks, with *POTUS* once again leading with the highest page rank score of 132.6. Some new additions such as *Coca Cola* and *Elon Musk* can be observed, the former being in the media for not pulling products out of Russia, and the latter being in opposition of President Putin's actions.

Moving on to the overall total scores of influence (Figure 12), the most influential users worldwide have been identified to be *ndtv* and *_theUt*, corresponding to an Indian news media company and an individual's account solely aimed at supporting Ukrainian efforts.

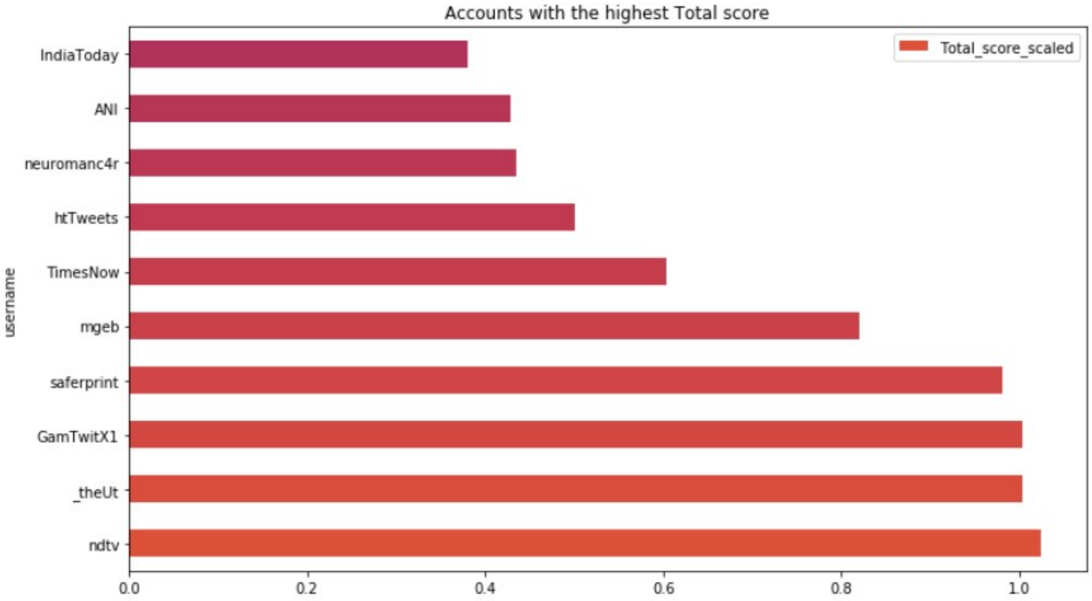


Figure 12: Most influential users globally

Looking at the most influential by country, top Ukrainian accounts (Figure 13) naturally include *_theUt*, as well as the official Twitter account of Ukraine and independent journalist *Olga Tokariuk*. Primary influencers in India (Figure 14) are overwhelmingly Indian news agencies, indicating that the perception of people might be affected by how the news is portrayed in India. In Non-NATO countries (Figure 15), top influencers are mostly news agencies from around the world such as *ndtv*, *ANI*, and *TimesNow*, whereas in countries such as the UK and the US (Figure 16 and Figure 17), most influential users seem to be individual people who have not ever previously been in the spotlight. This in fact is an indication of the power a single person can have, and that people globally and particularly in these countries, have all taken it upon themselves to voice their opinions and get involved, whether that is for or against the war.

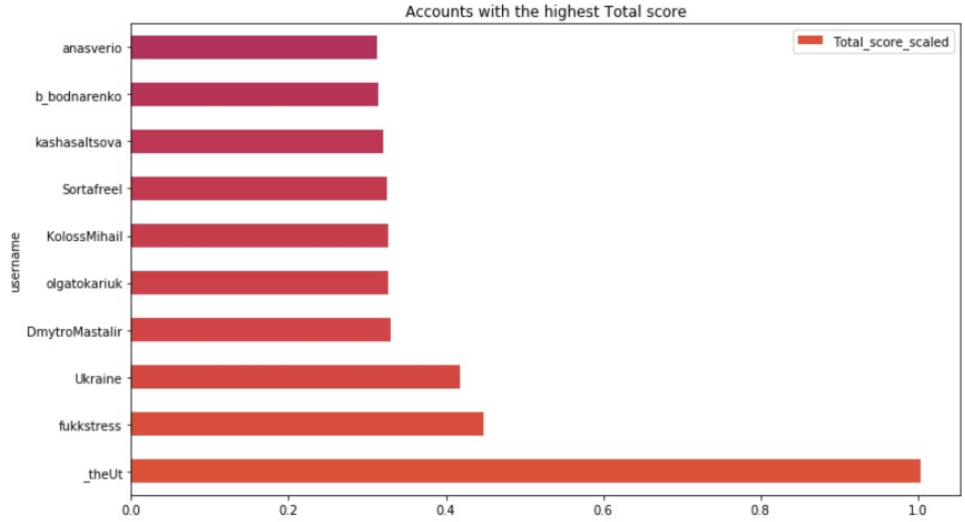


Figure 13: Most influential users in Ukraine

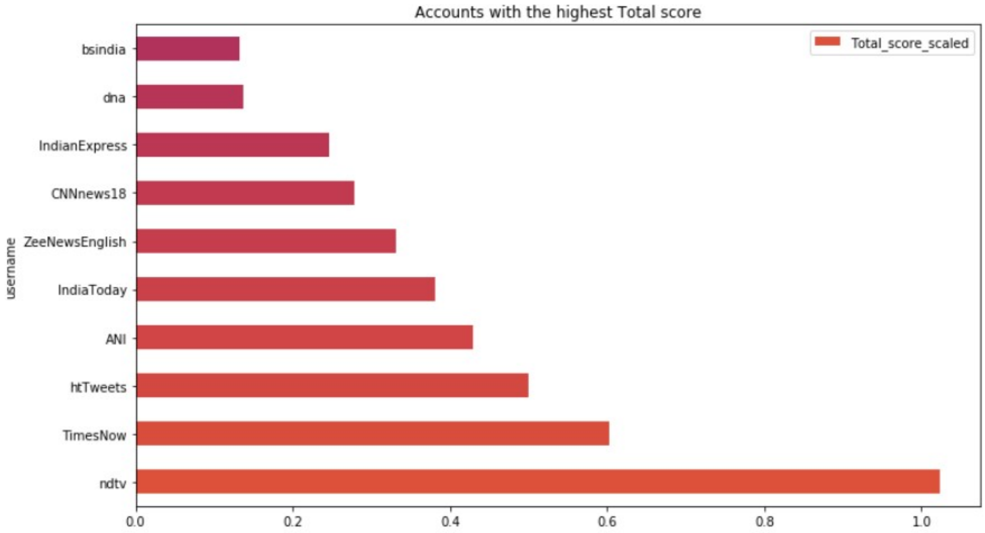


Figure 14: Most influential users in India

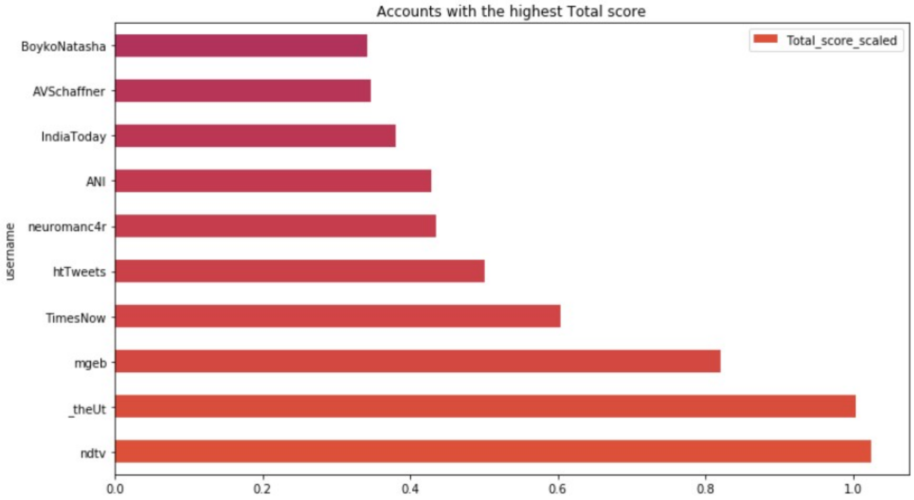


Figure 15: Most influential users in Non-NATO Countries

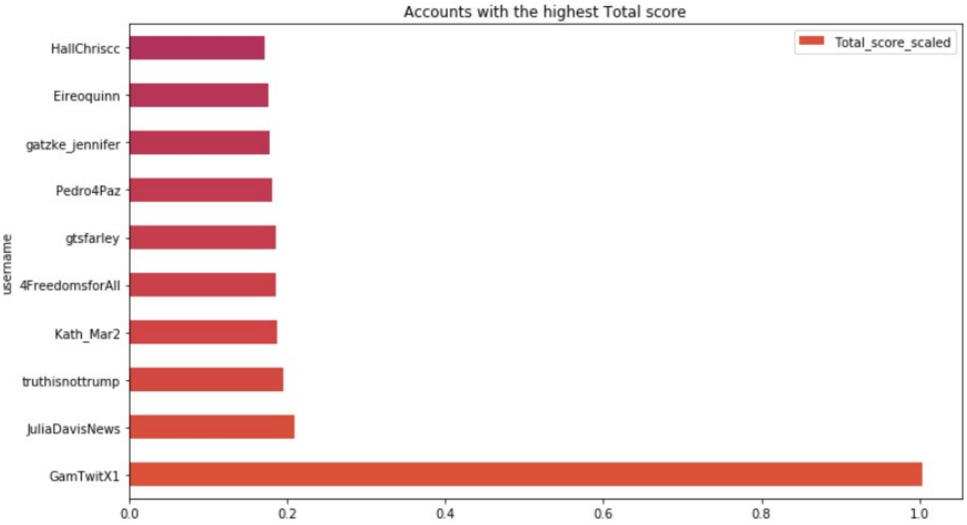


Figure 16: Most influential users in the US

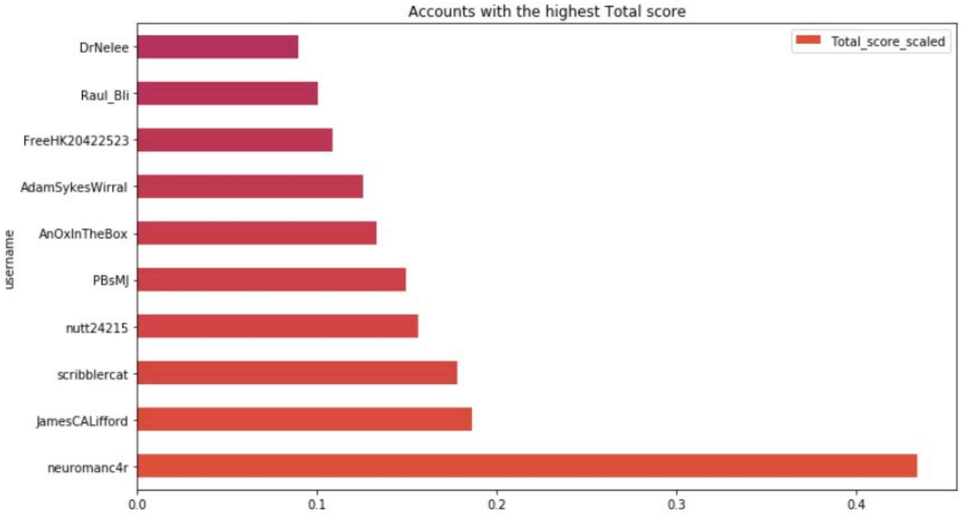
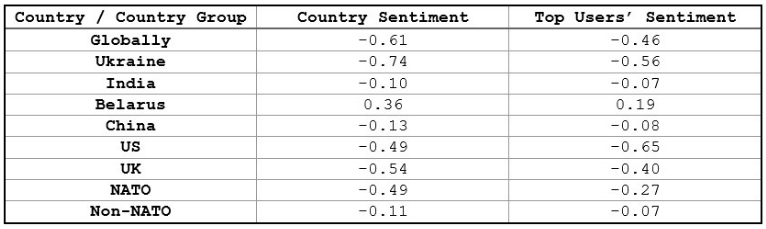


Figure 17: Most influential users in the UK

Having identified the key users of each country, we now proceed to analyse their sentiment. The following table indicates the sentiment of the users of each country or country group and the corresponding sentiment of each country/country group's most influential users.



Country / Country Group	Country Sentiment	Top Users' Sentiment
Globally	-0.61	-0.46
Ukraine	-0.74	-0.56
India	-0.10	-0.07
Belarus	0.36	0.19
China	-0.13	-0.08
US	-0.49	-0.65
UK	-0.54	-0.40
NATO	-0.49	-0.27
Non-NATO	-0.11	-0.07

Figure 18: Comparing top users' sentiment and overall sentiment in different countries/country groups

There seems to be a positive association between the sentiments, indicating that influencers share the views of the people they are representing. This perhaps is evidence of the persuasive powers of these users, or simply that the freedom of social media allows users who have views which are overwhelmingly in agreement with the views of the people, to come out of the shadows and represent the people as influencers and true "leaders".

It is also evident that for nearly all countries with the exception of the US, the absolute country sentiment is greater than that of the top user's sentiment. This might be as a result of top influential news agencies producing primarily tweets of a neutral sentiment, in order to not take a solid position on the war. It might also be due to the fact that people in the spotlight usually use more mild language and are perhaps are not as freely outspoken, compared to those who are not constantly in the public eye.

4.4 Streaming Processing Systems

Finally, in the last section we would like to design a system that retrieves news about the war in real-time. This is important as it might warn the population immediately of any possible future attacks and also keep them informed about any updates. Due to excessive amount of information available nowadays, its very complex to select the appropriate and accurate information. Twitter is a social media platform where people give their respective views on many topics, but it can also be a source of key information. Our task will be to identify the tweets related to the Russia-Ukraine conflict,obtain their subjectivity and polarity score, and finally filter the tweets that are ”neutral”. This is due to the fact that we would like the tweets to be as objective as possible and not obtain the ones that get carried away by emotion as they might not be the absolute truth. The subjectivity and polarity score are between $[0, 1]$ and $[-1, 1]$ respectively, hence a ”neutral” tweet will have a subjectivity score close to 0 *(0 being an objective tweet)* and a polarity close to 0 as well *(0 being a neutral sentiment tweet)* (Figure 19). Since there might be some error in the algorithm trying to predict the results and additionally it is hard to find tweets completely neutral (i.e tweets with subjectivity and polarity score equal to 0), an error is permitted. Specifically, an error of 10% for each of the scores. Hence, a ”neutral” tweet will have a subjectivity and polarity score between $[0, 0.1]$ and $[-0.1, 0.1]$, respectively.

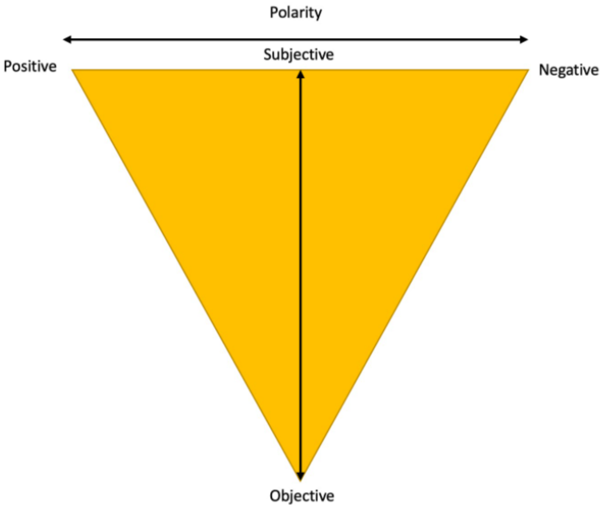


Figure 19: Visualizing Polarity and Subjectivity scores

4.4.1 Process

In order to obtain what we defined to be as neutral tweets, we will use the Twitter API to input tweets to the producer, which then will save and filter the tweets by the key words ”Russia”, ”Ukraine” and ”RussiaUkraineWar”. Then these tweets will be received by the consumer which will remove the symbols ””, ”RT”, ”:”, ”@+” and ”http§+”. Furthermore, the algorithm TextBlob will be applied to these processed tweets and the subjectivity and polarity scores will be obtained. In addition, the consumer will filter these tweets so that they have scores of subjectivity and polarity of $[0, 0.1]$ and $[-0.1, 0.1]$ respectively. Finally, the consumer will output the tweets satisfying the conditions in two ways: to the console in order to visualize the process and in a second instance to parquet files every 60 seconds by setting the trigger to be 60. The first process is run for 1000 batches, which takes approximately 6 minutes, and in order to maintain consistency, the second process will also be run for the same amount of time. Saving the results to a parquet file speeds up the process as these types of files are columnar and do not need much structure, which is beneficial in the streaming setting (Ahmed et al., 2017). In addition, when saving them to the parquet file we also specify checkpoints in order to have an end-to-end exactly-once fault tolerance (Armbrust et al., 2018). Storing the tweets can also be beneficial if the user prefers to read all the tweets together after a certain period of time instead of receiving news every second.

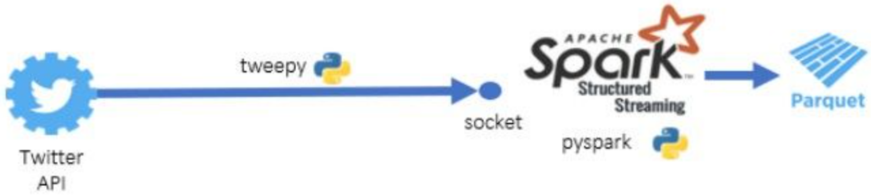


Figure 20: Twitter Streaming Pipeline

4.4.2 Results

The obtained results were surprising as the ”neutral” tweets in each batch were much fewer than expected. In many of the batches there were no tweets satisfying the imposed conditions. This may be due to the fact that this war has a great emotional charge and has led to the constant suffering of many individuals, hence it is hard for people not to take a position and maintain a neutral tone. Furthermore, it seems like the war has stopped being a trending topic recently as there are less tweets related to the topic. However, it is important to highlight that the designed system is not perfect, and although many of the tweets seem to have a neutral tone and provide useful information, there are some biased ones. These biased ones tend to come in the form of emojis, which are clearly recognized by reading the tweets, however the algorithm focuses on the words and misses completely the polarization and the subjectivity of the tweets. This implies that although in theory the system should give us the expected ”neutral” tweets, after testing it in practice the obtained results are not ideal. Nevertheless, the simplicity of the system being constructed makes it computationally efficient, which results in immediate outputs and permits the processing of large amount of data. These characteristics make it ideal for the streaming scenario and for our specific application of the system being a newsletter of the Russia-Ukraine conflict. To conclude, this system can be beneficial in other contexts aside from the one discussed in this paper and can be further improved to make it more reliable.

5. Conclusion

To summarize, this paper dealt with two scenarios. The first scenario dealt with the interpretation of the view of the vast majority of the world population as well as the noteworthy leaders with the help of Topic Modeling, Sentiment Analysis and Graph Network Analysis. After conducting these approaches, we concluded that the society as a whole is against the war. Further, the prominent leaders are also trying to do their best, in breaking this chain, by imposing sanctions on Russia. For instance, the United Kingdom has banned a number of trade deals with Russia, and released statements on their official UK Government website ”encouraging Russia to cease actions”. These efforts have been seen to be fueled by the voices of organisations, as well everyday ordinary people, who act as a driving force within the twitter community and have taken a leading role.

After conducting the discussed scenario which uncovered the desolation and despair caused by the war, led us to our second scenario, that comprises of developing a streaming process, that would act as an instant news tool of Russia-Ukraine war developments for the entire globe. One cannot disregard the power of Twitter and the vast amounts of information that may be uncovered simply by human interactions within the platform. Tapping into this limitless source of data and uncovering the truth within, one could get notified through the use of certain phrases and words, about the upcoming incidents, attacks and evacuation plans instantaneously. This, thus serves as an alternative and perhaps more lucrative approach to traditional news sources, given that the news is coming directly from the people, and who better to tell a story than the ones living it.

It is paramount, to note that our study had some weaknesses. The first one being that there was no twitter data from Russia. The reason for the same is that the Russian state had restricted access to all social media websites such as Twitter and Facebook (Bonifacic). Another, essential weakness was reducing our dataset to the tweets that were made only in English, as English is the most widely spoken language and is easier to understand from the perspective of people from around the world (Zhang, 2013).

Today, marks the 62^{nd} day of the war and the situation is still not under-control despite people on all sides being against the war. Some reasonable ways in which we can improve our analysis is analysing tweets of languages other than English, which could also give a more representative view of the sentiment of people around the world and which could provide further insight into the development of the war. Moreover, obtaining Russian data from other social media platforms, would enable analysis of their mind-set and would provide a better understanding of the thought processes of the Russian population. This would further enhance our news streaming model, as individuals from other territories will be intimated about what the plans and road-map of the country at the epicentre of this war are. The streaming system could also be improved to handle the problem of obtaining biased tweets by using more complex algorithms that are able to handle irregular text and emoticons such as VADER. One must however bear in mind, that the time complexity of the algorithm remains low, given that are our priority is to retrieve the desired tweets without delays. One can only hope however, that peace prevails and that this newsletter system is used in more positive manners; not in the context of a war.

References

Shahzad Ahmed, M Usman Ali, Javed Ferzund, Muhammad Atif Sarwar, Abbas Rehman, and Atif Mehmood. Modern data formats for big bioinformatics data analytics. *arXiv preprint arXiv:1707.05364*, 2017.

Wajdi Aljedaani, Furqan Rustam, Stephanie Ludi, Ali Ouni, and Mohamed Wiem Mkaouer. Learning sentiment analysis for accessibility user reviews. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 239–246. IEEE, 2021.

Thomas Ambrosio. The political success of russia-belarus relations: Insulating minsk from a color revolution. *Demokratizatsiya*, 14(3), 2006.

Michael Armbrust, Tathagata Das, Joseph Torres, Burak Yavuz, Shixiong Zhu, Reynold Xin, Ali Ghodsi, Ion Stoica, and Matei Zaharia. Structured streaming: A declarative api for real-time applications in apache spark. In *Proceedings of the 2018 International Conference on Management of Data*, pages 601–613, 2018.

Steven Bird, Ewan Klein, and Edward Loper. Nltk documentation. *Online: accessed April*, 2008.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Bonifacic. Russia restricts twitter access amid ukraine invasion. URL `https://www.engadget.com/russia-restricts-domestic-twitter-access-172821677.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAIzPSWslMt4ExBc7QpfsArDhyL5tDXIPYg29xl3ulLNOB56BjRJAJwKOm82-z0CknKnKFbqlcqXKGccuQnM-_kxMLJMmVPUbPzO4kjT8T8u_dDJnVXAgmmYCTwuVXQAk_EsSLo0rfB3ybYUc3AOE2DGQwrnSnu99tU_mSRjL9dCx`.

Venkateswarlu Bonta and Nandhini Kumaresh2and N Janardhan. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6, 2019.

Christine M. Kowalczyk Daniel N. Jones Brian J. Taillon, Steven M. Mueller. Understanding the relationships between social media influencers and their followers: the moderating role of closeness. 2020.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 10–17, 2010.

S Chakravarthy. Tokenization for natural language processing, 2020.

Bill Chambers and Matei Zaharia. *Spark: The definitive guide: Big data processing made simple.* " O'Reilly Media, Inc.", 2018.

Abhishek Akshay Chaudhri, SS Saranya, and Sparsh Dubey. Implementation paper on analyzing covid-19 vaccines on twitter dataset using tweepy and text blob. *Annals of the Romanian Society for Cell Biology*, pages 8393–8396, 2021.

Sanket Chintapalli, Derek Dagit, Bobby Evans, Reza Farivar, Thomas Graves, Mark Holderbaugh, Zhuo Liu, Kyle Nusbaum, Kishorkumar Patil, Boyang Jerry Peng, et al. Benchmarking streaming computation engines: Storm, flink and spark streaming. In *2016 IEEE international parallel and distributed processing symposium workshops (IPDPSW)*, pages 1789–1792. IEEE, 2016.

Carlos Omar Cortés Hinojosa. Probabilistic topic modeling with latent dirichlet allocation on apache spark. 2016.

Ankur Dave, Alekh Jindal, Li Erran Li, Reynold Xin, Joseph Gonzalez, and Matei Zaharia. Graphframes: an integrated api for mixing graph and relational queries. In *Proceedings of the fourth international workshop on graph data management experiences and systems*, pages 1–8, 2016.

Mingxing Duan, Kenli Li, Zhuo Tang, Guoqing Xiao, and Keqin Li. Selection and replacement algorithms for memory performance improvement in spark. *Concurrency and Computation: Practice and Experience*, 28(8): 2473–2486, 2016.

Shihab Elbagir and Jing Yang. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, page 16, 2019.

Nada Elgendy and Ahmed Elragal. Big data analytics: a literature review paper. In *Industrial conference on data mining*, pages 214–227. Springer, 2014.

Hossam Elzayady, Khaled M Badran, and Gouda I Salama. Sentiment analysis on twitter data using apache spark framework. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pages 171–176. IEEE, 2018.

Massimo Franceschet. Pagerank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6): 92–101, 2011.

M Govindarajan. Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. *International Journal of Advanced Computer Research*, 3(4):139, 2013.

J Praveen Gujjar and HR Prasanna Kumar. Sentiment analysis: Textblob for decision making. *International Journal of Scientific Research & Engineering Trends*, (7):1097–1099, 2021.

Shelley Gupta, Archana Singh, and Jayanthi Ranjan. Emoji score and polarity evaluation using cldr short name and expression sentiment. In *International Conference on Soft Computing and Pattern Recognition*, pages 1009–1016. Springer, 2020.

Nazarii Gutsul and Kristina Khrul. *Multicultural Societies and their Threats: Real, hybrid and media wars in Eastern and South-Eastern Europe*, volume 10. LIT Verlag Münster, 2017.

I Hemalatha, GP Saradhi Varma, and A Govardhan. Sentiment analysis tool using machine learning algorithms. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(2):105–109, 2013.

Geoffrey A Hosking. *Russia: people and empire, 1552-1917*. Harvard University Press, 1997.

Ling Huang, Jinyu Ma, and Chunling Chen. Topic detection from microblogs using t-lda and perplexity. In *2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW)*, pages 71–77. IEEE, 2017.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78 (11):15169–15211, 2019.

RE Johnson. The road to unfreedom: Russia, europe, america by timothy snyder, 2020.

Ruqaiya Khanam and Abhishek Sharma. Sentiment analysis using different machine learning techniques for product review. In *2021 International Conference on Computational Performance Evaluation (ComPE)*, pages 646–650. IEEE, 2021.

Christopher SG Khoo and Sathik Basha Johnkhan. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511, 2018.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

Steven Loria et al. textblob documentation. *Release 0.15*, 2:269, 2018.

Manogna Meduru, Antara Mahimkar, Krishna Subramanian, Puja Y Padiya, and Prathmesh N Gunjgur. Opinion mining using twitter feeds for political analysis. *Int. J. Comput.(IJC)*, 25(1):116–123, 2017.

Ismael Solis Moreno, Peter Garraghan, Paul Townend, and Jie Xu. An approach for characterizing workloads in google cloud to derive realistic resource utilization models. In *2013 IEEE Seventh International Symposium on Service-Oriented System Engineering*, pages 49–60. IEEE, 2013.

Lekha R Nair and DR Sujala D Shetty. Streaming twitter data analysis using spark for effective job search. *Journal of Theoretical & Applied Information Technology*, 80(2), 2015.

Edi Surya Negara, Dendi Triadi, and Ria Andryani. Topic modelling twitter data with latent dirichlet allocation method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 386–390. IEEE, 2019.

Heather Newman and David Joyner. Sentiment analysis of student evaluations of teaching. In *International conference on artificial intelligence in education*, pages 246–250. Springer, 2018.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Evgenia Papageorgiou. A predictive model for customer satisfaction. 2021.

Sigit Priyanta, I Nyoman Prayana Trisna, and Nyoman Prayana. Social network analysis of twitter to identify issuer of topic using pagerank. *International Journal of Advanced Computer Science and Applications*, 10(1): 107–111, 2019.

Sergio Ramírez-Gallego, Alberto Fernández, Salvador García, Min Chen, and Francisco Herrera. Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce. *Information Fusion*, 42:51–61, 2018.

Paul Robinson. Russia's role in the war in donbass, and the threat to european security. *European Politics and Society*, 17(4):506–521, 2016.

Semih Sahin. *Memory optimizations for distributed executors in big data clouds.* PhD thesis, Georgia Institute of Technology, 2019.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

Pedro Saleiro and Carlos Soares. Learning from the news: Predicting entity popularity on twitter, 2016.

Jonathan Samosir, Maria Indrawan-Santiago, and Pari Delir Haghighi. An evaluation of data stream processing systems for data driven applications. *Procedia Computer Science*, 80:439–449, 2016.

Theresa Schmiedel, Oliver Müller, and Jan vom Brocke. Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, 22(4):941–968, 2019.

Anton Shakhov et al. User voice overview: topic recognition and sentiment analysis of customer feedback in the b2c sector. 2020.

Chad Sharp, Jelle van Assema, Brian Yu, Kareem Zidane, and David J Malan. An open-source, api-based framework for assessing the correctness of code in cs50. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, pages 487–492, 2020.

Bhupender Singh Shekhawat. *Sentiment Classification of Current Public Opinion on BREXIT: Naïve Bayes Classifier Model vs Python's TextBlob Approach.* PhD thesis, Dublin, National College of Ireland, 2019.

Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. Gervader-a german adaptation of the vader sentiment analysis tool for social media texts. In *LWDA*, pages 178–189, 2019.

Andrew T Wolff. The future of nato enlargement after the ukraine crisis. *International Affairs*, 91(5):1103–1121, 2015.

Babak Yadranjiaghdam, Seyedfaraz Yasrobi, and Nasseh Tabrizi. Developing a real-time data analytics framework for twitter streaming data. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 329–336. IEEE, 2017.

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 15–28, 2012.

Nashwan Dheyaa Zaki, Nada Yousif Hashim, Yasmin Makki Mohialden, Mostafa Abdulghafoor Mohammed, Tole Sutikno, and Ahmed Hussein Ali. A real-time big data sentiment analysis for iraqi tweets using spark streaming. *Bulletin of Electrical Engineering and Informatics*, 9(4):1411–1419, 2020.

Bei Zhang. An analysis of spoken language and written language and how they affect english language learning and teaching. *Journal of Language Teaching & Research*, 4(4), 2013.

Tatiana Zhurzhenko et al. "capital of despair". holodomor memory and political conflicts in kharkiv after the orange revolution. *East European Politics and Societies*, 25(03):597–639, 2011.

6. Statement of Title

This project had four sections and, we hereby, confirm that we all have contributed, equally to all the sections of this project.

- 1. Topic Modeling
- 2. Sentiment Analysis
- 3. Graph Network Analysis
- 4. Streaming processing systems